Homework 9 (Due at 11:59pm on 2015/03/27)

## Description

Write a program to use different string similarity functions to match a string to its closest resemblance. You are given two restaurant datasets **d1.txt** and **d2.txt**. Each dataset has 4 fields: **name**, **address**, **city** and **type**. Please use all string similarity functions in List 1 to match the values for each field. An example for matching the name field is shown below. For each field (name, address, city and type), your program should output a table. The columns are all the similarities in List 1. It shows the most similar counterparts in d2.txt for each value in d1.txt (all the rows). If there is a tie for multiple values in d2.txt, use the first one from these values.

You can use the library listed in the reference to do the homework. You can also use any third-party tool or your own code to calculate the string similarities if they can correctly calculate the similarities in List 1.

**Measure** the results generated by different string similarity functions using **Accuracy**. To calculate this metric, you can find the ground truth in the **groundtruth.csv** file. It has pairs of matching records. It shows the matching value for each value of all the fields. You program should compare the results with the ground truth to find the number of true positive matching and then compute the accuracy by dividing the true positive by the number of records in d1.txt.

**Example of names from two data files:**

|  |
| --- |
| **d1.txt** |
| Cafe bizou |
| le chardonnay |
| … |

|  |
| --- |
| **d2.txt** |
| Bizou Cafe |
| chardonnay |
| Cafe bizz |
| … |

**Example of expected Output:**

|  |  |  |  |
| --- | --- | --- | --- |
| **d1** | Levenshtein | TF\*IDF | … |
| Cafe bizou | Cafe bizz | bizou Cafe |  |
| le chardonnay | chardonnay | chardonnay |  |
| … | … | … |  |

**List 1: String similarity function:**

* Levenshtein distance
* Needleman-wunch
* Smith-waterman
* JaroWinkler
* TF/IDF
* Soundex

**References:**

[1] string similarity calculating functions: <http://code.google.com/p/java-similarities/source/browse/trunk/simmetrics/src/main/java/uk/ac/shef/wit/simmetrics/>

[2] accuracy <http://en.wikipedia.org/wiki/Precision_and_recall>

## Submission

Please turn in your homework on Blackboard with the following files, which should be named as **hw9\_[firstname]\_[lastname].zip**

(1) A report that contains the following information:

* the output table as shown in the example for each field
* the accuracy for each similarity
* compare the performance of different similarities and explain the differences.

(2) Your program

(3) A readme file explaining

* how to run your code
* all the libraries you used in your code